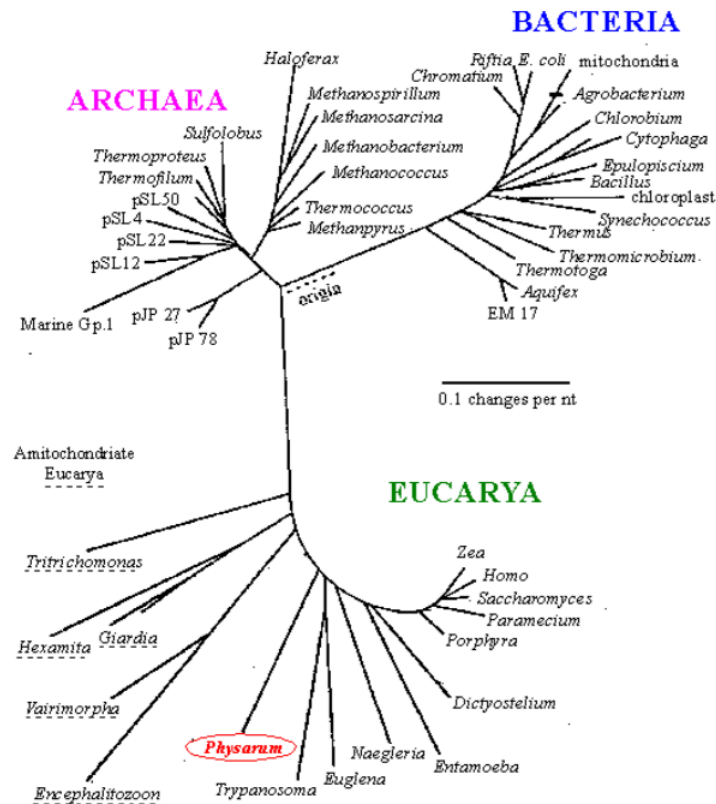


# Pairwise Alignment



# Sequences are related

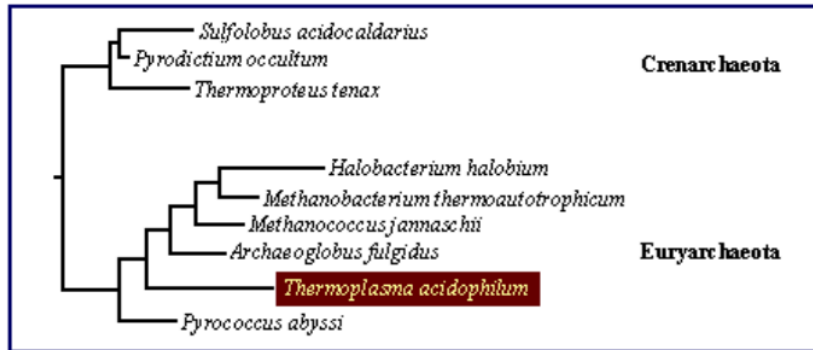
- Darwin: all organisms are related through descent with modification
- => Sequences are related through descent with modification
- => Similar molecules have similar functions in different organisms



Phylogenetic tree based on  
ribosomal RNA:  
three domains of life

# Why compare sequences?

---



- Determination of evolutionary relationships

Protein 1: binds oxygen



Sequence similarity

Protein 2: binds oxygen ?

- Prediction of protein function and structure (database searches).

# Pairwise alignments

---

43.2% identity;

Global alignment score: 374

```

          10          20          30          40          50
alpha  V-LSPADKTNVKA AWGKVGAHAGEYGA EALERMFLSFPTTKTYFPHF-DLS-----HGSA
       : :: :. : : : : :. : : : : : :. : : :. : : : : :. :
beta   VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
          10          20          30          40          50

          60          70          80          90         100         110
alpha  QVKGHGKKVADALTNAVAHVDDMPNALSASDLHAHKL RVPVNFKLLSHCLLVTLAAHL
       ..... : ..... : ..... : ..... : ..... :. :. :.
beta   KVKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHF
       60          70          80          90         100         110

          120         130         140
alpha  PAEFTPAVHASLDKFLASVSTVLTSKYR
       : : : : :. : : : : :. :.
beta   GKEFTPPVQAAYQKVVAGVANALAHKYH
       120         130         140
```

# Pairwise alignments

---

43.2% identity;

Global alignment score: 374

```

              10      20      30      40              50
alpha  V-LSPADKTNVKA AWGKVG AHAGEYGA EALERMFLSFPTTKTYFPHF-DLS-----HGSA
        :  ::  ::  :  :  :::  ..  :  :::  :::  ..  :  :  :  :  :::  :
beta   VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
              10      20      30      40      50

              60      70      80      90      100     110
alpha  QVKGHGKKVADALTNAVAHVDDMPNALSASDLHAHKL RVPVNFKLLSHCLLVTLAAHL
        .....:  .....:  .....:  .....:  .....:  ..  :  :
beta   KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRL LGNVLVCVLAHHF
        60      70      80      90      100     110

              120     130     140
alpha  PAEFTPAVHASLDKFLASVSTVLTSKYR
        :::  :::  :  .....:
beta   GKEFTPPVQAAYQKVVAGVANALAHKYH
        120     130     140
```

100.000% identity in 3 aa overlap

SPA  
:::  
SPA

# Alignment scores: match vs. mismatch

---

Simple scoring scheme (too simple in fact...):

Matching amino acids: 5

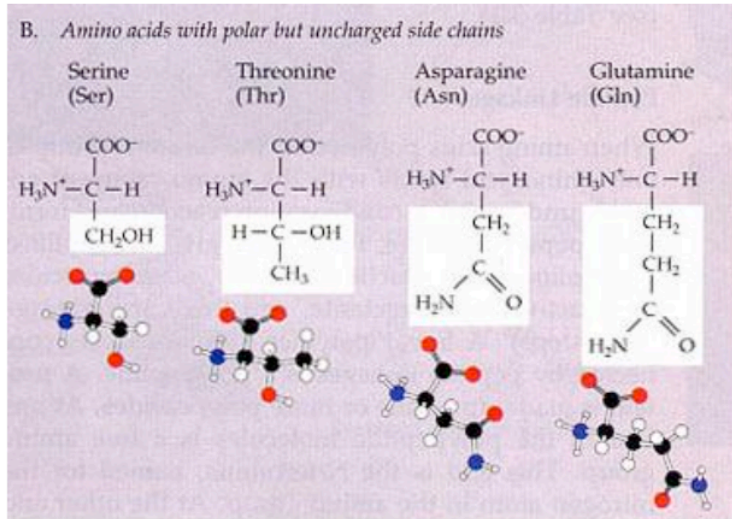
Mismatch: 0

Scoring example:

K	A	W	S	A	D	V	
:	:	:	:	:	:	:	
K	D	W	S	A	E	V	

$$5+0+5+5+5+0+5 = 25$$

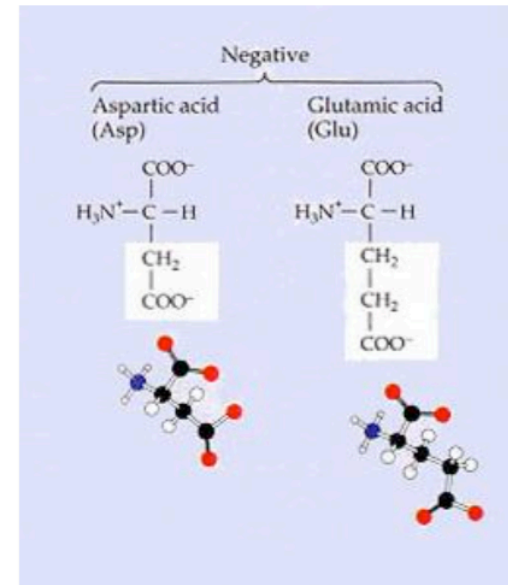
# Amino acid properties



Serine (S) and Threonine (T) have similar physicochemical properties

=> Substitution of S/T or E/D occurs relatively often during evolution

=> Substitution of S/T or E/D should result in scores that are only moderately lower than identities



Aspartic acid (D) and Glutamic acid (E) have similar properties

# Pairwise alignments: conservative substitutions

---

43.2% identity;

Global alignment score: 374

```

      10      20      30      40      50
alpha  V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
      : : . : . : : : : : : : : : : : : : : : : : : : : : : :
beta   VHLTPPEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
      10      20      30      40      50

      60      70      80      90     100     110
alpha  QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
      . : . : . : . : . : . : . : . : . : . : . : . : . : . : . :
beta   KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHF
      60      70      80      90     100     110

      120     130     140
alpha  PAEFTPAVHASLDKFLASVSTVLTSKYR
      : : : : : : : : : : : : : :
beta   GKEFTPPVQAAYQKVVAGVANALAHKYH
      120     130     140
```



# Protein substitution matrices

## BLOSUM50 matrix:

- Positive scores on diagonal (identities)
- Similar residues get higher (positive) scores
- Dissimilar residues get smaller (negative) scores

<b>A</b>	<b>5</b>																			
<b>R</b>	-2	<b>7</b>																		
<b>N</b>	-1	-1	<b>7</b>																	
<b>D</b>	-2	-2	<b>2</b>	<b>8</b>																
<b>C</b>	-1	-4	-2	-4	<b>13</b>															
<b>Q</b>	-1	<b>1</b>	0	0	-3	<b>7</b>														
<b>E</b>	-1	0	0	<b>2</b>	-3	<b>2</b>	<b>6</b>													
<b>G</b>	0	-3	0	-1	-3	-2	-3	<b>8</b>												
<b>H</b>	-2	0	<b>1</b>	-1	-3	<b>1</b>	0	-2	<b>10</b>											
<b>I</b>	-1	-4	-3	-4	-2	-3	-4	-4	-4	<b>5</b>										
<b>L</b>	-2	-3	-4	-4	-2	-2	-3	-4	-3	<b>2</b>	<b>5</b>									
<b>K</b>	-1	<b>3</b>	0	-1	-3	<b>2</b>	<b>1</b>	-2	0	-3	-3	<b>6</b>								
<b>M</b>	-1	-2	-2	-4	-2	0	-2	-3	-1	<b>2</b>	<b>3</b>	-2	<b>7</b>							
<b>F</b>	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	<b>1</b>	-4	0	<b>8</b>						
<b>P</b>	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	<b>10</b>					
<b>S</b>	<b>1</b>	-1	<b>1</b>	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	<b>5</b>				
<b>T</b>	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	<b>2</b>	<b>5</b>			
<b>W</b>	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	<b>1</b>	-4	-4	-3	<b>15</b>		
<b>Y</b>	-2	-1	-2	-3	-3	-1	-2	-3	<b>2</b>	-1	-1	-2	0	<b>4</b>	-3	-2	-2	<b>2</b>	<b>8</b>	
<b>V</b>	0	-3	-3	-4	-1	-3	-3	-4	-4	<b>4</b>	<b>1</b>	-3	<b>1</b>	-1	-3	-2	0	-3	-1	<b>5</b>
	<b>A</b>	<b>R</b>	<b>N</b>	<b>D</b>	<b>C</b>	<b>Q</b>	<b>E</b>	<b>G</b>	<b>H</b>	<b>I</b>	<b>L</b>	<b>K</b>	<b>M</b>	<b>F</b>	<b>P</b>	<b>S</b>	<b>T</b>	<b>W</b>	<b>Y</b>	<b>V</b>

# Pairwise alignments: insertions/deletions

---

43.2% identity;

Global alignment score: 374

```

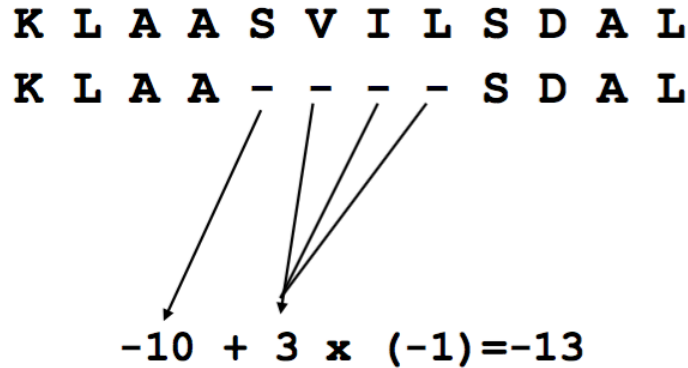
                10         20         30         40                     50
alpha  V-LSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHF-DLS----HGSA
      :  ::  ::  :  :  ::::  ..  :  :::::  ....  :  :  :  :  :::  :.
beta   VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
                10         20         30         40         50

                60         70         80         90         100        110
alpha  QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
      .....:  .....:  .....:  .....:  .....:  .....:  .....:  :.
beta   KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHF
                60         70         80         90         100        110

                120        130        140
alpha  PAEFTPAVHASLDKFLASVSTVLTSKYR
      ::::  ::::  :.  .....:  :..
beta   GKEFTPPVQAAYQKVVAGVANALAHKYH
      120        130        140
```

# Alignment scores: insertions/deletions

---



Affine gap penalties:

Multiple insertions/deletions may be one evolutionary event =>

Separate penalties for **gap opening** and **gap elongation**

# Handout

---

Compute 4 alignment scores: two different alignments using two different alignment matrices (and the same gap penalty system)

Score 1: Alignment 1 + BLOSUM-50 matrix + gaps

Score 2: Alignment 1 + BLOSUM-Trp matrix + gaps

Score 3: Alignment 2 + BLOSUM-50 matrix + gaps

Score 4: Alignment 2 + BLOSUM-Trp matrix + gaps



Note: fake matrix constructed for pedagogic purposes.

## Handout: summary of results

---

	Alignment 1	Alignment 2
BLOSUM-50	38	51
BLOSUM-Trp	118	91

# Protein substitution matrices: different types

---

- **Identity matrix**  
(match vs. mismatch)
- **Chemical properties matrix**  
(use knowledge of physicochemical properties to design matrix)
- ➔ • **Empirical matrices**  
(based on observed pair-frequencies in hand-made alignments)
  - PAM series
  - BLOSUM series
  - Gonnet

Searching for the optimal  
alignment...

# The problem:

## How many possible alignments are there?

---

ACG	AC-G	--ACG	-A-CG
ACG	ACG-	AC-G-	A-CG-
-ACG	AC-G	--ACG	...
ACG-	A-CG	A-CG-	
-ACG	AC-G	--ACG	
AC-G	-ACG	AC--G	
-ACG	ACG-	--ACG	
A-CG	AC-G	A-C-G	
A-CG	ACG-	--ACG	
ACG-	A-CG	A--CG	
A-CG	ACG-	-A-CG	
AC-G	-ACG	ACG--	
A-CG	--ACG	-A-CG	
-ACG	ACG--	AC-G-	



# The problem: How many possible alignments are there?

---

ACG	AC-G	--ACG	-A-CG
ACG	ACG-	AC-G-	A-CG-
-ACG	AC-G	--ACG	...

Two protein sequences of length 100 amino acids can be aligned in approximately  $10^{60}$  different ways

Time needed to test all possibilities is same order of magnitude as the entire lifetime of the universe.

A-CG	ACG-	--ACG
ACG-	A-CG	A--CG
A-CG	ACG-	-A-CG
AC-G	-ACG	ACG--
A-CG	--ACG	-A-CG
-ACG	ACG--	AC-G-

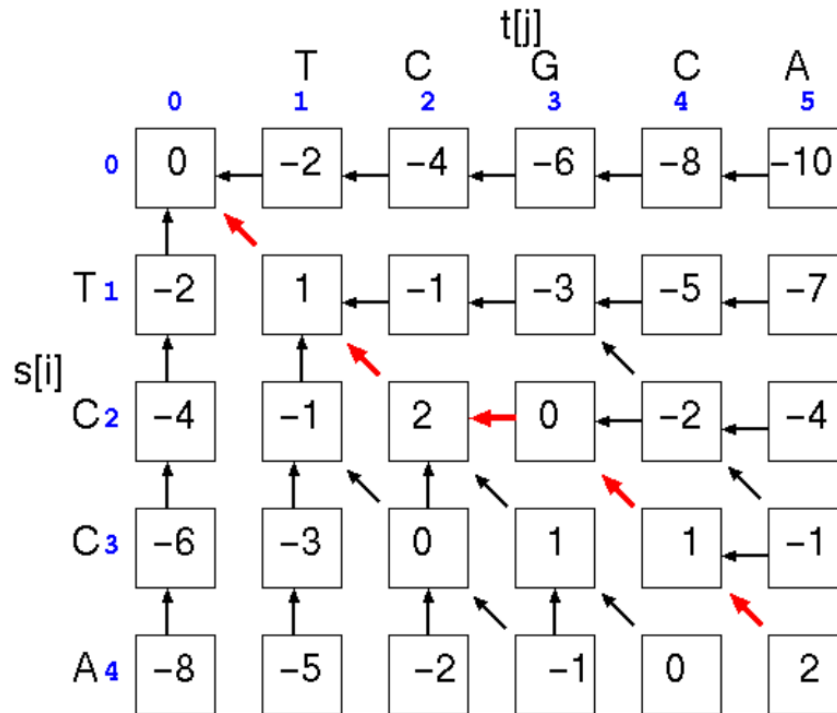
Solution:  
Dynamic programming

TCGCA

TCCA

# Pairwise alignment: the solution

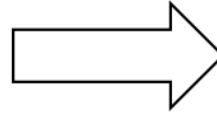
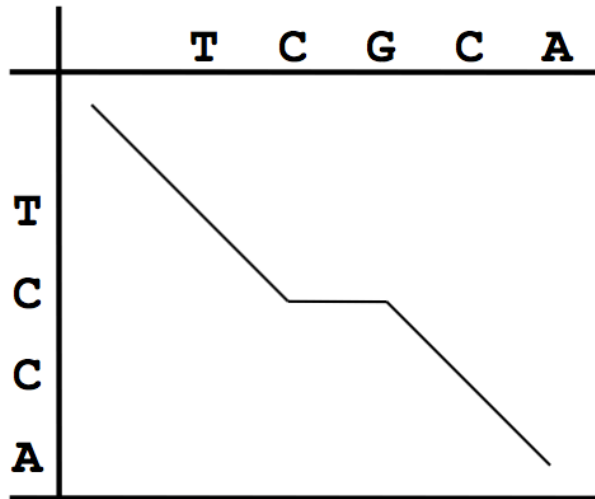
**"Dynamic programming"**  
(the Needleman-Wunsch algorithm)



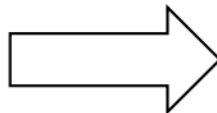
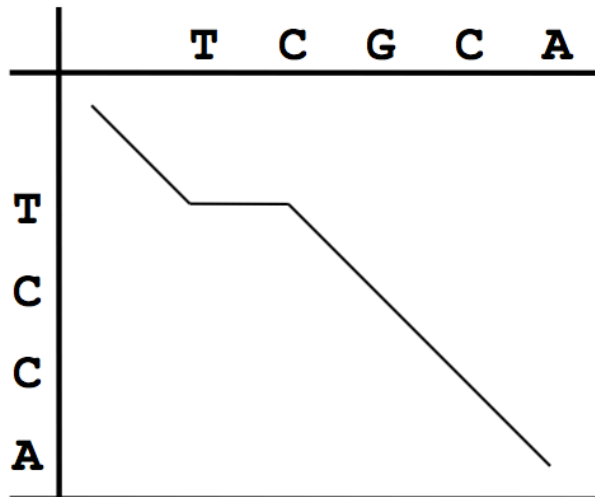
Filling a scoring  
matrix

# Alignment depicted as path in matrix

---



**TCGCA**  
**TC-CA**



**TCGCA**  
**T-CCA**

# Dynamic programming: computation of scores

---

	T	C	G	C	A
T					
C					
C					
A					

Any given point in matrix can only be reached from three possible previous positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

# Dynamic programming: computation of scores

---

	T	C	G	C	A
T					
C		x			
C					
A					

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

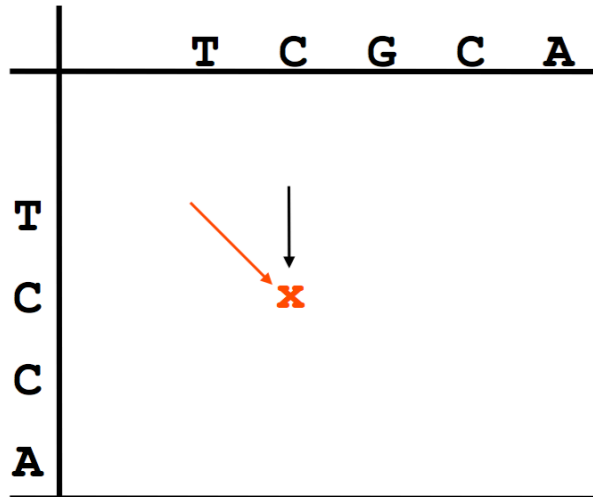
=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

$$\text{score}(x,y) = \max \left\{ \begin{array}{l} \text{score}(x,y-1) - \text{gap-penalty} \\ \text{score}(x-1,y) \\ \text{score}(x-1,y-1) \end{array} \right.$$

# Dynamic programming: computation of scores

---

	T	C	G	C	A
T					
C					
C					
A					



Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.


$$\text{score}(x,y) = \max \left\{ \begin{array}{l} \text{score}(x,y-1) - \text{gap-penalty} \\ \text{score}(x-1,y-1) + \text{substitution-score}(x,y) \end{array} \right.$$



# Dynamic programming: computation of scores

---

	T	C	G	C	A
T					
C					
C					
A					



Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

$$\text{score}(x,y) = \max \begin{cases} \text{score}(x,y-1) - \text{gap-penalty} \\ \text{score}(x-1,y-1) + \text{substitution-score}(x,y) \\ \text{score}(x-1,y) - \text{gap-penalty} \end{cases}$$

# Dynamic programming: computation of scores

---

	T	C	G	C	A
T					
C					
C					
A					

Any given point in matrix can only be reached from three possible positions (you cannot “align backwards”).

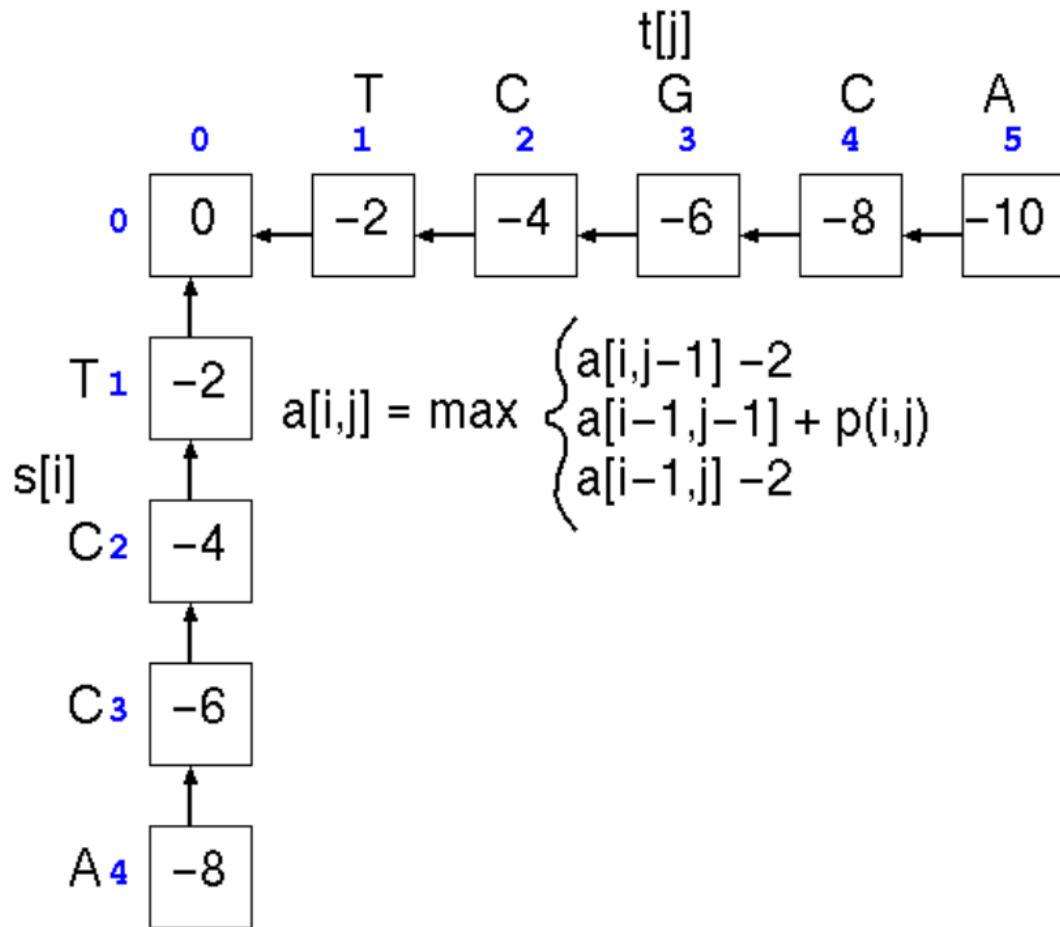
=> Best scoring alignment ending in any given point in the matrix can be found by choosing the highest scoring of the three possibilities.

Each new score is found by choosing the maximum of three possibilities.  
For each square in matrix: keep track of where best score came from.

Fill in scores one row at a time, starting in upper left corner of matrix, ending in lower right corner.

$$\text{score}(x,y) = \max \begin{cases} \text{score}(x,y-1) - \text{gap-penalty} \\ \text{score}(x-1,y-1) + \text{substitution-score}(x,y) \\ \text{score}(x-1,y) - \text{gap-penalty} \end{cases}$$

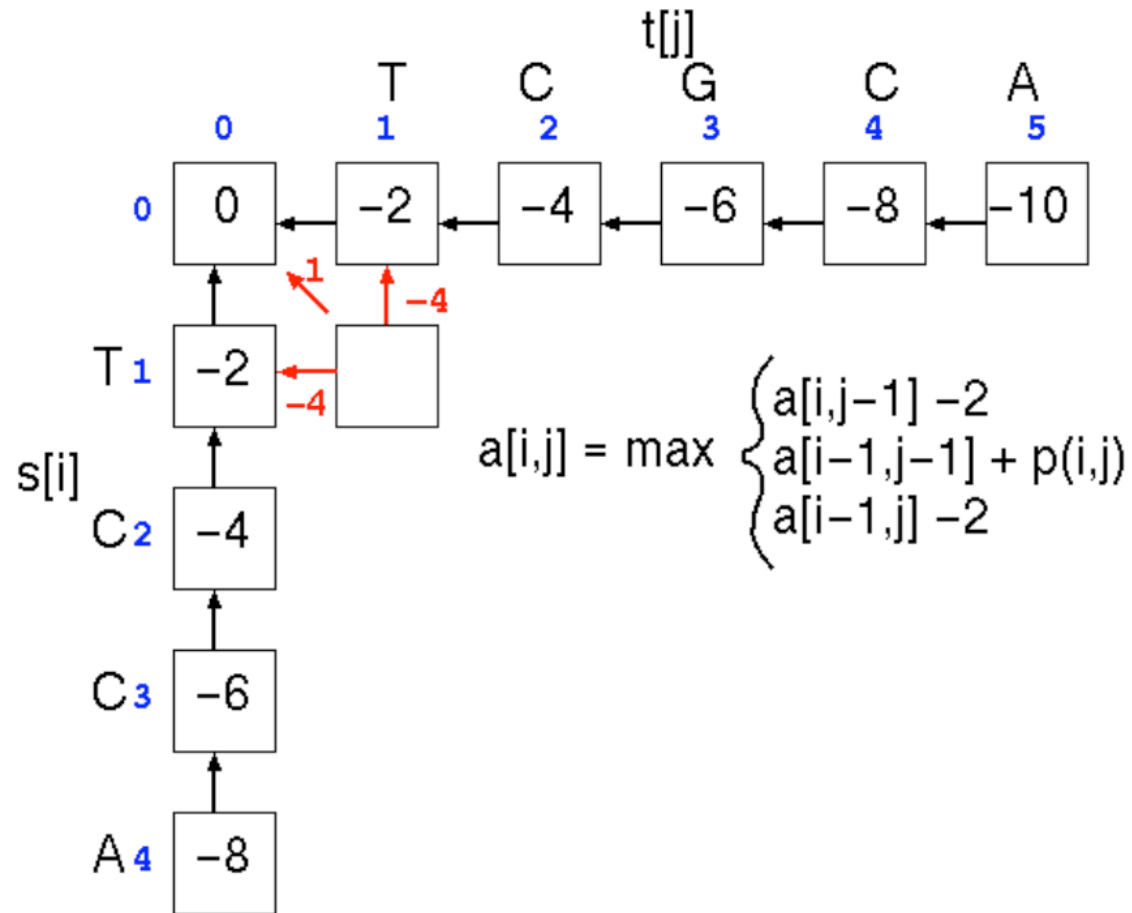
# Dynamic programming: example



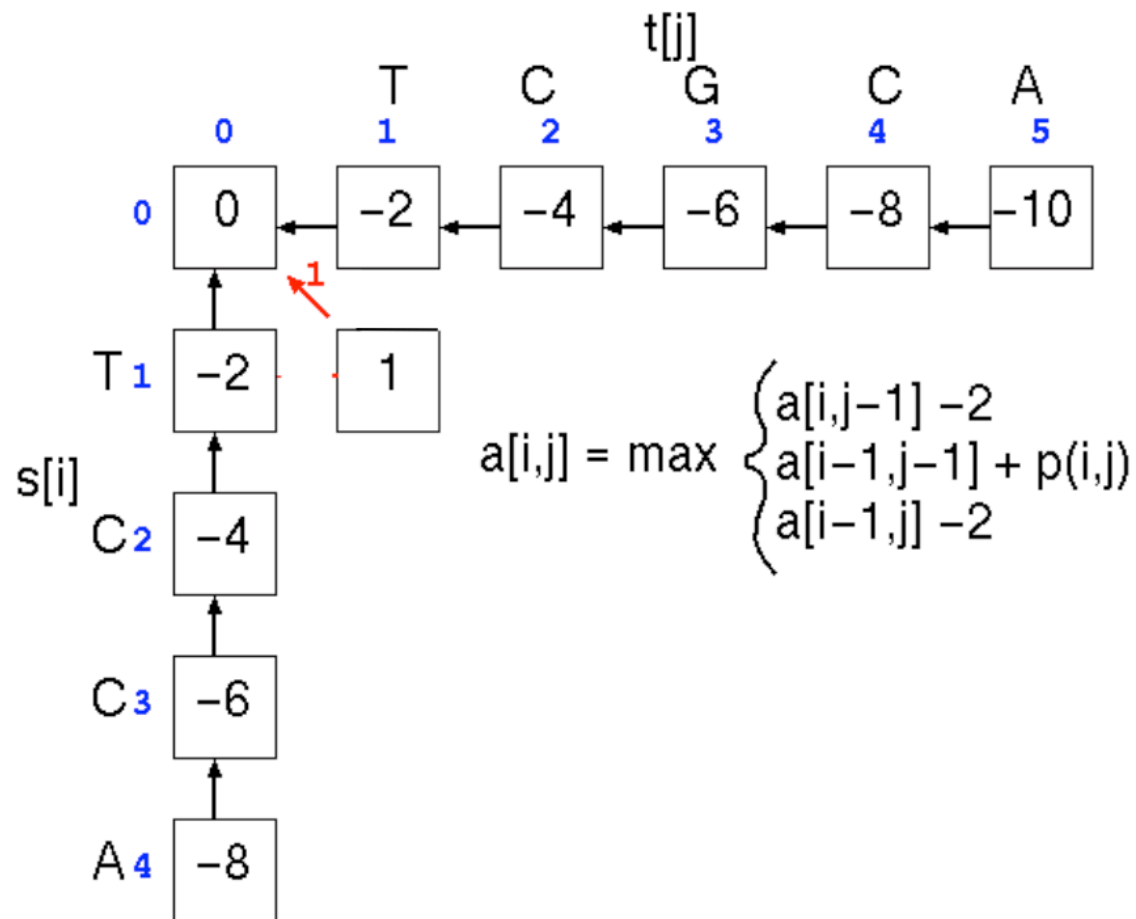
	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Gaps: -2

# Dynamic programming: example

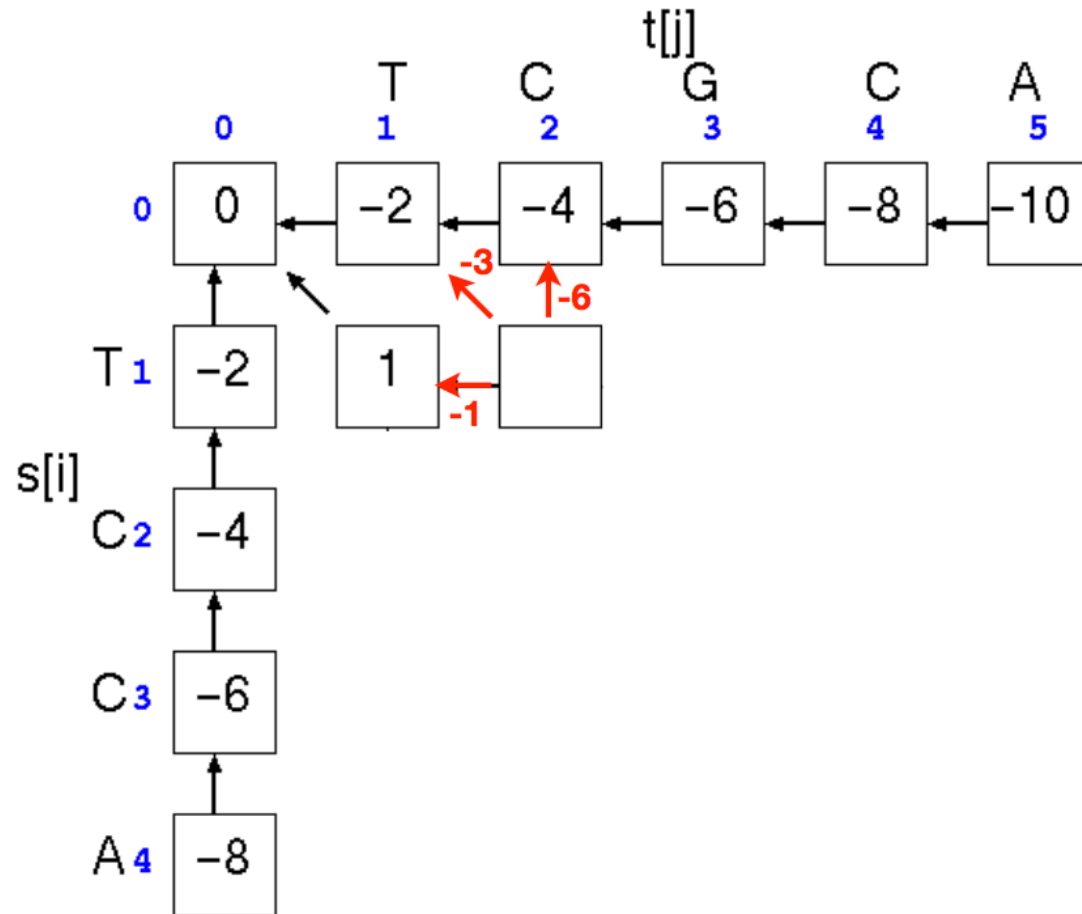


# Dynamic programming: example



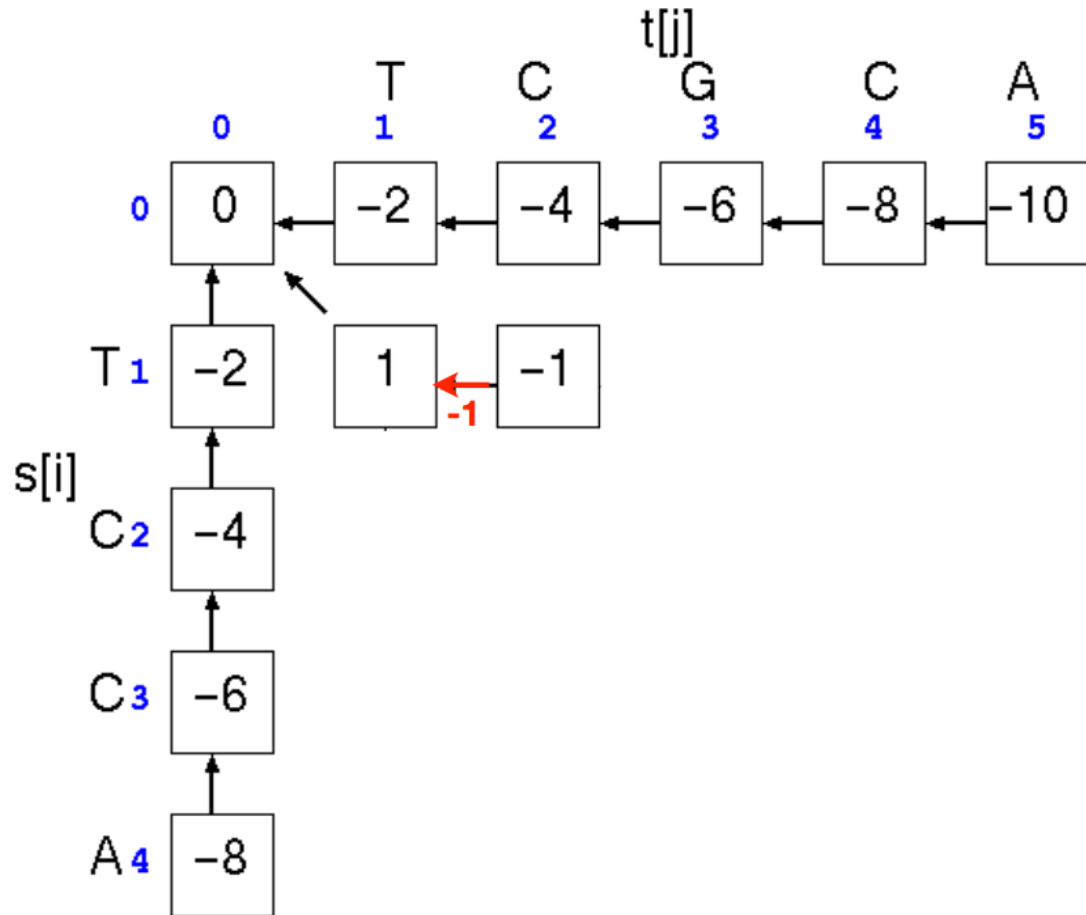
# Dynamic programming: example

---

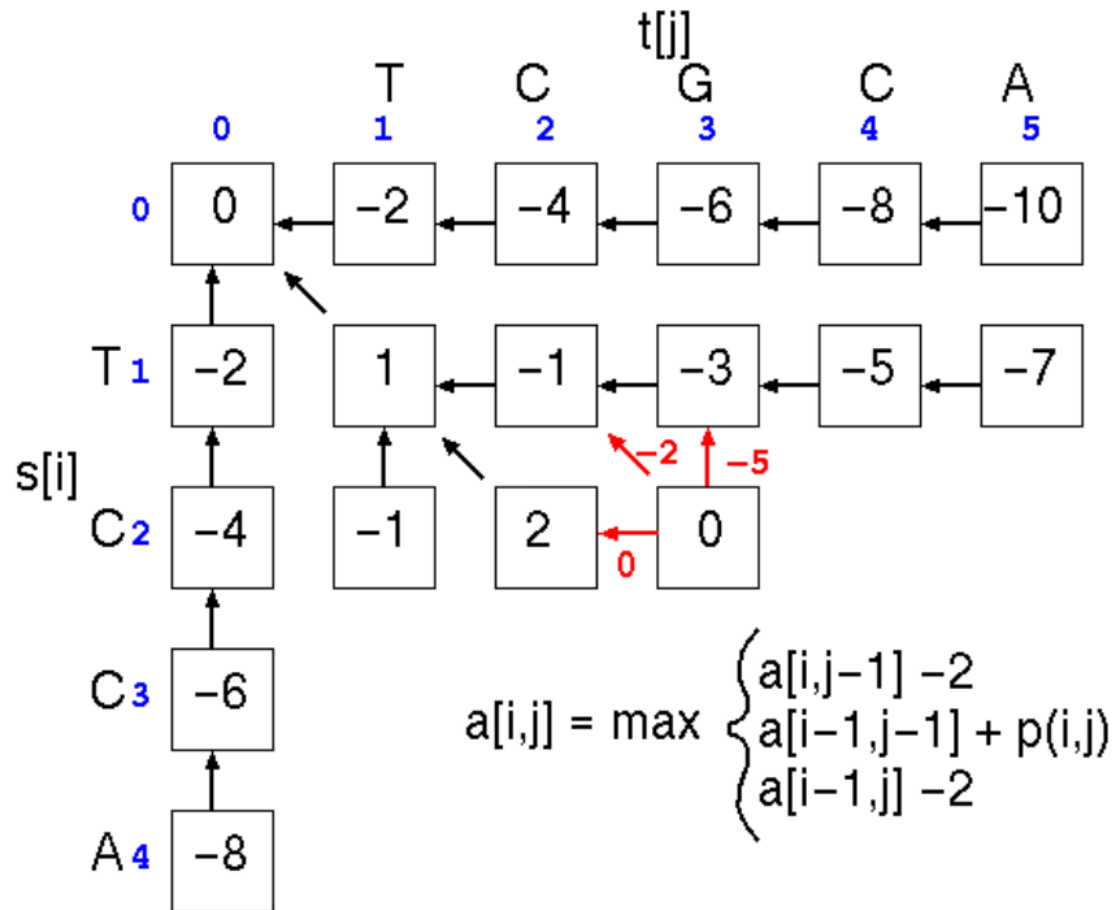


# Dynamic programming: example

---

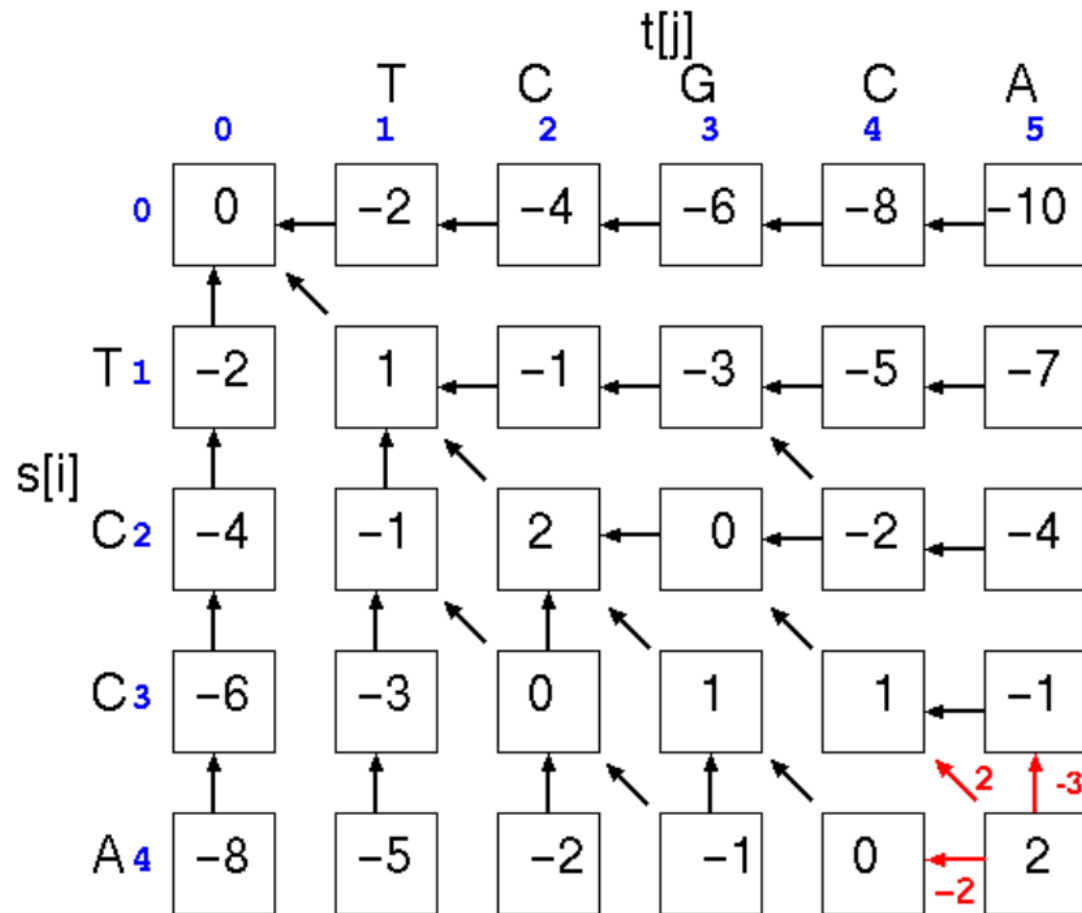


# Dynamic programming: example

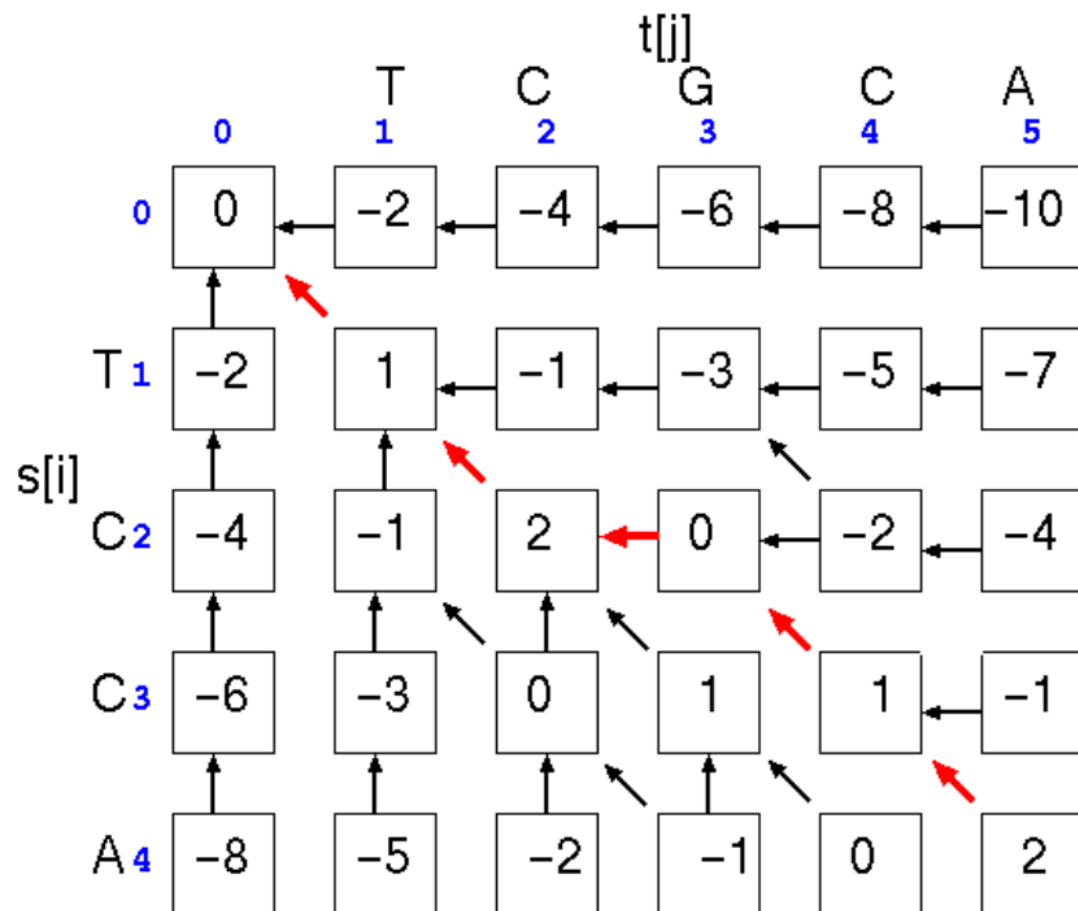




# Dynamic programming: example



# Dynamic programming: example



$$\begin{array}{c}
 \text{T} \quad \text{C} \quad \text{G} \quad \text{C} \quad \text{A} \\
 \vdots \quad \vdots \quad \quad \vdots \quad \vdots \\
 \text{T} \quad \text{C} \quad - \quad \text{C} \quad \text{A} \\
 \hline
 1+1-2+1+1 = \underline{2}
 \end{array}$$

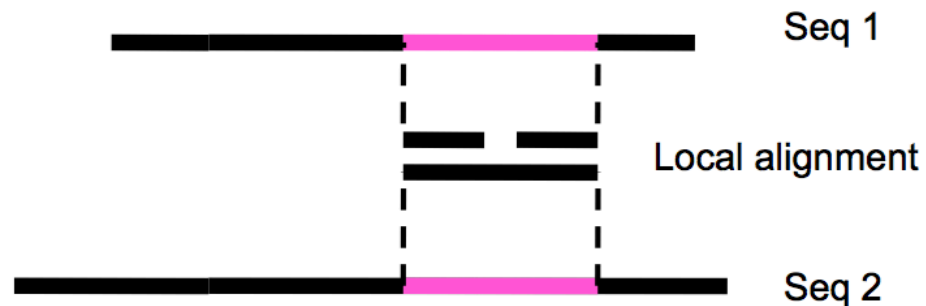
# Global versus local alignments

---

Global alignment: align full length of both sequences.  
(The “Needleman-Wunsch” algorithm).



Local alignment: find best partial alignment of two sequences  
(the “Smith-Waterman” algorithm).



# Local alignment overview

---

- The recursive formula is changed by adding a fourth possibility: zero. This means local alignment scores are never negative.

$$\text{score}(x,y) = \max \begin{cases} \text{score}(x,y-1) - \text{gap-penalty} \\ \text{score}(x-1,y-1) + \text{substitution-score}(x,y) \\ \text{score}(x-1,y) - \text{gap-penalty} \\ 0 \end{cases}$$

- Trace-back is started at the highest value rather than in lower right corner
- Trace-back is stopped as soon as a zero is encountered

# Local alignment: example

		H	E	A	G	A	W	G	H	E	E
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	2	0	<b>20</b>	<b>12</b>	4	0	0
H	0	10	2	0	0	0	12	18	<b>22</b>	14	6
E	0	2	16	8	0	0	4	10	18	<b>28</b>	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

AWGHE

AW-HE

# Substitution matrices and sequence similarity

---

- Substitution matrices come as series of matrices calculated for different degrees of sequence similarity (different evolutionary distances).
- "Hard" matrices are designed for similar sequences
  - Hard matrices are designated by high numbers in the BLOSUM series (e.g., BLOSUM80)
- "Soft" matrices are designed for less similar sequences
  - Soft matrices have low BLOSUM values (45)

# Substitution matrices and sequence similarity

---

- Substitution matrices come as series of matrices calculated for different degrees of sequence similarity (different evolutionary distances).
- "Hard" matrices are designed for similar sequences
  - Hard matrices are designated by high numbers in the BLOSUM series (e.g., BLOSUM80)
  - Hard matrices yield short, highly conserved alignments
- "Soft" matrices are designed for less similar sequences
  - Soft matrices have low BLOSUM values (45)
  - Soft matrices yield longer, less well conserved alignments

# What did you learn?

- Purpose of sequence alignment:
  - Find **evolutionary** relationships
  - Predict protein **function**
- The optimal alignment can be found using dynamic programming:
  - **Local** alignment: Smith-Waterman
  - **Global** alignment: Needleman-Wunsch
- Optimal alignment means having the **best possible score** given:
  1. Substitution matrix
  2. Set of gap penalties
- The optimal alignment is NOT necessarily the most biologically meaningful



# Exercise 1 – Pairwise Alignment

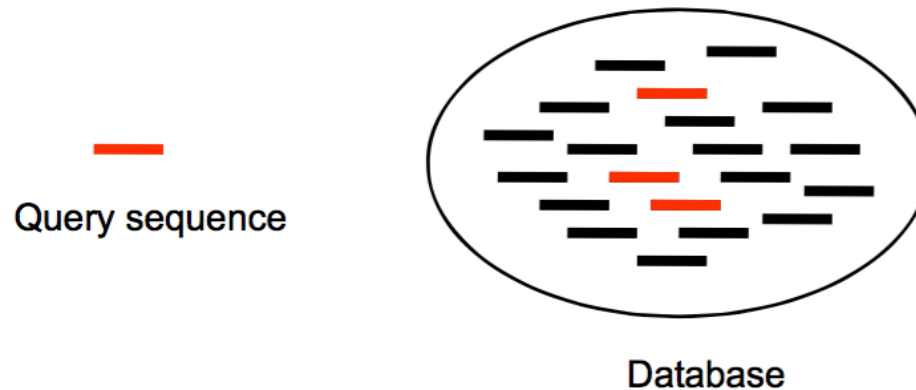
# Database Searching



# Database searching

---

**Using pairwise alignments to search  
databases for similar sequences**



## Database searching

---

Most common use of pairwise sequence alignments is to search databases for related sequences. For instance: find probable function of newly isolated protein by identifying similar proteins with known function.

Most often, **local alignment** ( “Smith-Waterman”) is used for database searching: you are interested in finding out if ANY domain in your protein looks like something that is known.

Often, full Smith-Waterman is too time-consuming for searching large databases, so heuristic methods are used (fasta, BLAST).

## Database searching: heuristic search algorithms

---

### FASTA (Pearson 1995)

Uses heuristics to avoid calculating the full dynamic programming matrix

Speed up searches by **an order of magnitude** compared to full Smith-Waterman

The statistical side of FASTA is still stronger than BLAST

### BLAST (Altschul 1990, 1997)

Uses rapid word lookup methods to completely skip most of the database entries

**Extremely fast**

One order of magnitude faster than FASTA

Two orders of magnitude faster than Smith-Waterman

Almost as sensitive as FASTA

# BLAST flavors

---

## BLASTN

Nucleotide query sequence  
Nucleotide database

## BLASTP

Protein query sequence  
Protein database

## BLASTX

Nucleotide query sequence  
Protein database  
Compares all six reading frames with  
the database

## TBLASTN

Protein query sequence  
Nucleotide database  
"On the fly" six frame translation of  
database

## TBLASTX

Nucleotide query sequence  
Nucleotide database  
Compares all reading frames of query  
with all reading frames of the  
database

# Searching on the web: BLAST at NCBI

Very fast computers dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface

*But you still need knowledge about BLAST to use it properly*

The screenshot displays the NCBI BLAST web interface in a browser window. The address bar shows the URL: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&Q=blast+ncbi>. The page title is "Protein BLAST: search protein databases using a protein query". The interface includes a navigation bar with "Home", "Recent Results", "Saved Strategies", and "Help". A "My NCBI" link is also present. The main form is titled "Enter Query Sequence" and contains a large text input field for the query sequence, a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. Below this is a section for uploading a file, with a "Choose File" button and a "Job Title" input field. The "Choose Search Set" section includes a "Database" dropdown menu set to "Non-redundant protein sequences (nr)", an "Organism" input field, and an "Entrez Query" input field. The "Program Selection" section shows the "blastp (protein-protein BLAST)" algorithm selected. At the bottom, there is a "BLAST" button and a checkbox for "Show results in a new window". The footer contains copyright information and links to "Disclaimer", "Privacy", "Accessibility", "Contact", and "Send feedback on new interface".

# When is a database hit significant?

---

- **Problem:**

- Even **unrelated** sequences can be aligned (yielding a low score)
- How do we know if a database hit is **meaningful**?
- When is an **alignment score** sufficiently high?

- **Solution:**

- Determine the range of alignment scores you would expect to get for **random reasons** (i.e., when aligning unrelated sequences).
- Compare actual scores to the **distribution of random scores**.
- Is the real score much higher than you'd **expect by chance**?

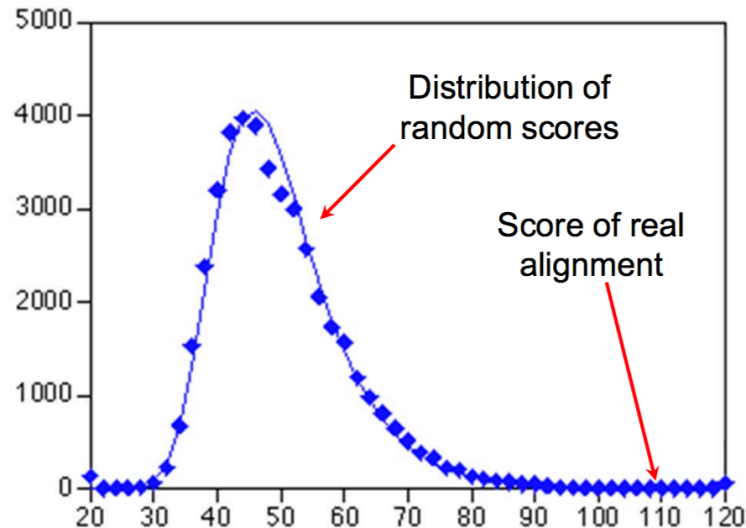


## Significance of alignment score expressed as E-value

---

Searching a database of **unrelated** sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the exact nature of the database and the query sequence



**E-value:** the number of **random hits** to **expect** for any given score

Want E-values below 1 (the lower the better)

# What did you learn?

- Heuristic methods (e.g. BLAST) is faster to search databases than Smith-Waterman
- BLASTN, BLASTP etc.
- E-value: Measure for significance of database hit

## Exercise 2 – Using BLAST